Europe
An Interdisciplinary Journal of Research, Literature and Literary Reviews
Vol. 104 No. 1, 2026 ISSN: 0014-2751

GHOST IN THE ARCHIVE: USING COMPUTATIONAL STYLOMETRY TO RECOVER LOST VOICES IN DIGITIZED MARGINALIZED LITERATURE

Ilyar Kuzminsk
Staatliche Akademie der Bildenden Künste Stuttgart
Germany

Abstract

Traditional literary archives are structural artifacts of power, historical contingencies, and systematic biases, frequently erasing or silencing marginalized voices. While large-scale digitization efforts have successfully preserved vast physical corpora, simply converting paper to pixel does not solve the underlying exclusions of the archival record. Unattributed, pseudonymous, or collectively published prose often hides the idiosyncratic hand of minor writers, women, queer creators, and racialized individuals within dominant, mainstream structures. This review paper synthesizes the burgeoning cross-disciplinary field of computational stylometry as a radical method of archival recovery. By leveraging algorithmic pattern recognition—including lexical, syntactic, character n-gram, and advanced deep learning vector frameworks—scholars can extract unique stylistic markers to identify, classify, and re-attribute prose written by suppressed or anonymous authors within large unmapped datasets. Framed within structural digital humanities, poststructuralist archival theory, and the political economy of data, this article maps the technical workflow of algorithmic text attribution, analyzes its primary methodological and mathematical underpinnings, critically reviews established case studies, and outlines the deep ethical boundaries governing machine-assisted text reassignment. Ultimately, we argue that computational stylometry serves as an indispensable hermeneutic intervention, resurrecting the 'ghosts in the archive' to permanently correct and democratize the literary canon.

Keywords: Keywords: Computational Stylometry, Archival Recovery, Digital Humanities, Authorship Attribution, Marginalized Literature, Machine Learning, Postcolonial Archives.

1. Introduction and Thematic Imperative

The institutional archive has never been a passive, neutral repository of objective human history. Rather, as poststructuralist and postcolonial theorists have long argued, the archive is a highly policed apparatus of state power, cultural hegemony, and canonical validation. The historical record is fundamentally structured by absences, erasures, and systematic violences that determine which texts are carefully cataloged and preserved, and which are discarded, suppressed, or relegated to permanent anonymity. Marginalized writers—encompassing women, queer individuals, colonized subjects, racial minorities, and working-class agitators—have historically been blocked from accessing overt networks of literary production and distribution. To escape state censorship, social ostracization, or economic ruin,

these authors frequently resorted to pseudonyms, published anonymously, or permitted their distinct voices to be subsumed into collective cultural vehicles or corporate broadsides.

In the twenty-first century, mass digitization campaigns driven by institutional libraries, national consortia, and corporate entities (such as Google Books, the Internet Archive, and HathiTrust) have dramatically shifted the material realities of literary study. Millions of historical records, periodicals, novels, and pamphlets have been transitioned into digital spaces, ostensibly democratizing access to human textuality. However, this transition exposes a profound epistemological gap: the simple conversion of physical documents to optical character recognized (OCR) text does not fix the underlying exclusions of the catalog. If a marginalized author's prose is trapped within an anonymous corporate file or falsely attributed to a dominant male editor, basic digital access fails to resolve the historical injustice. Digitization without advanced, intelligent algorithmic remediation merely recreates and hardens the patriarchal and colonial biases of the traditional physical archive into high-speed digital architectures.

This review paper focuses on this exact methodological and political crossroads. We examine the capacity of computational stylometry—the quantitative study of linguistic style through statistical and computational pattern recognition—to act as an assertive tool of archival recovery. While traditional computational stylometry has been applied to high-profile canonical mysteries (such as verifying the Shakespearean or Marlovian canons, or unmasking contemporary pseudonyms like J.K. Rowling), its most profound, transformative future lies in recovering suppressed, subaltern, and minor literary voices. By shifting computational focus away from celebrated canonical authors toward the vast, unmapped landscape of marginalized and non-attributed prose, digital humanities can operationalize stylistic signatures as keys to unlock historical identity.

> *Archival Recovery Axiom: Computational stylometry transforms text from a static container of historic information into an active multi-dimensional vector space. In doing so, it exposes the subtle, sub-conscious linguistic micro-patterns that historical editors, state censors, and canonical gatekeepers could not consciously alter or erase.*

To establish a comprehensive, high-impact framework for this paradigm, this review systematically covers four critical axes: First, we contextualize computational stylometry within poststructuralist archival theory, mapping the shift from Close Reading to Distant Reading and examining the data ethics of structural recovery. Second, we offer a granular analysis of the computational and mathematical architectures of stylometric attribution, moving from low-level feature extraction (lexical, syntactic, and character-level n-grams) to advanced machine learning and deep learning classification paradigms. Third, we provide a cross-disciplinary synthesis of breakthrough case studies where computational tools have successfully restored agency to historically erased authors. Finally, we map the severe

structural challenges facing this methodology—including low-quality OCR datasets, lack of training data, stylistic evolution, and the deep ethical risks of computational colonization.

2. Theoretical Framework: Archival Silences and Digital Interventions

2.1 Poststructuralist Archival Theory and the Politics of Erasure

To understand the role of algorithmic recovery, we must first destabilize the concept of the archive itself. In his foundational text Archive Fever, Jacques Derrida argued that the archive is not merely a collection of past documents, but the literal locus where political power is institutionalized and maintained. The power of the archive resides in its gatekeeping capacity: the 'archivization' process dictates what can be thought, spoken, or remembered. Michel Foucault similarly theorized the archive as the general system of the formation and transformation of statements, operating as a structural regulatory mechanism that links knowledge to power. When a dominant socio-political class controls the archive, the historical experiences and artistic productions of marginalized communities are not simply ignored; they are systematically filtered out through deliberate omissions, strategic misattributions, and calculated material destruction.

Michel-Rolph Trouillot, in Silencing the Past, demonstrated that historical erasure occurs at multiple critical junctures: during the moment of fact creation (the making of the source), fact assembly (the making of the archive), fact retrieval (the making of the narrative), and retrospective significance. For marginalized authors, the silence begins at the moment of production. A nineteenth-century Black woman writing an abolitionist critique might be forced to publish her essay under the name of her white employer or within an anonymous collective newspaper column to ensure safety and distribution. Consequently, the traditional archivist, relying strictly on metadata and explicitly written title pages, codifies these historical injustices into stable, permanent library catalogs. Standard archival practice, bound to institutional parameters of provenance and explicit attribution, is fundamentally ill-equipped to undo erasures that occurred before the text ever reached the archivist’s desk.

2.2 From Close Reading to Distant Reading: The Hermeneutic Shift

For generations, literary scholarship responded to archival silences through intensive close reading, archival excavation, and biographical reconstruction. This hermeneutic mode, while invaluable, suffers from severe limits of scale and visibility. Close reading requires the human scholar to engage deeply with a singular text, relying on conscious aesthetic intuition to spot historical anomalies or stylistic variations. Yet, when marginalized voices are buried within hundreds of thousands of pages of anonymous, digitized, multi-authored nineteenth-century radical periodicals, close reading is limited by human cognitive capacity. A researcher cannot read a century of daily newspapers to find an elusive voice.

This limitation prompted Franco Moretti to propose 'distant reading,' a methodology that bypasses the intensive study of individual texts in favor of algorithmic exploration across vast textual corpora. Distant reading allows the literary scholar to focus on global structures, macro-level thematic movements, and statistical configurations across thousands of works simultaneously. Matthew Jockers extended this paradigm into 'macroanalysis,' demonstrating

that computational text analysis reveals underlying patterns of genre, theme, and stylistic evolution that remain entirely invisible to the naked human eye. In recovering marginalized literature, computational stylometry operationalizes distant reading not as a method to ignore the micro-text, but as a critical macro-filter. It scans immense, anonymized digital fields to highlight specific, mathematically anomalous clusters that warrant closer human critical attention.

2.3 The Epistemology of Data Recovery and Algorithmic Reparation

When applied to marginalized literature, computational stylometry shifts from a purely technical classification tool to an assertive form of algorithmic reparation. This methodology rejects the idea that a text's metadata is fixed or unchangeable. Instead, it views the text as a dynamic, layered site of computational evidence. The core epistemological premise is that an author's writing style is driven by stable, unconscious linguistic habits—such as the frequency of highly common function words, specialized punctuation rhythms, and characteristic character transitions. These sub-cognitive patterns operate beneath the conscious control of the writer and far beyond the editing strokes of a censor. Consequently, even when an author's name is completely stripped from a document, their computational 'fingerprint' remains embedded within the text's structural profile. Using machine learning models to detect these hidden markers allows digital humanities scholars to construct a counter-archive that challenges, expands, and corrects the structural biases of traditional histories.

3. Methodological and Mathematical Foundations of Computational Stylometry

The application of computational stylometry to historical archives requires a rigorous, multi-tiered pipeline of technical preparation, feature extraction, and machine learning classification. This section outlines the operational architecture required to transform messy, historical digital archives into mathematically sound environments capable of high-accuracy authorship attribution.

3.1 Text Preprocessing, OCR Rectification, and Tokenization

Historical digital archives present unique computational challenges due to the degraded quality of source materials and the limitations of historical printing. Texts derived from nineteenth-century or early twentieth-century printings often suffer from severe OCR degradation, including broken characters, systemic spelling variations, and misread punctuation. Before executing stylometric feature extraction, a rigorous preprocessing pipeline is mandatory. This begins with programmatic OCR error correction, utilizing deep-learning contextual models (such as specialized BERT variants) trained on historical corpora to restore degraded tokens without destroying individual stylistic spelling traits.

Following error correction, texts undergo tokenization, where continuous text strings are split into discrete linguistic units (words, punctuation, morphosyllables). Unlike standard NLP tasks which remove non-alphanumeric entities and smooth variations, stylometric preprocessing must handle punctuation and spelling choices with care, as these elements

often carry significant stylistic signal. Lemmatization and Part-of-Speech (POS) tagging are then executed using historical language models to ensure that archaic grammatical structures are accurately identified and retained for syntactic feature engineering.

3.2 The Stylometric Feature Hierarchy

To map an author's stylistic signature, computational tools extract features across a multi-layered linguistic hierarchy. These features vary in structural abstraction and computational complexity, moving from simple word counts to deep syntactic configurations:

Lexical Features and Most Frequent Words (MFWs): 1. Lexical Features and Most Frequent Words (MFWs): The bedrock of classic and contemporary stylometry rests on the relative frequency distribution of Most Frequent Words, primarily function words (prepositions, conjunctions, articles, pronouns). While content words reflect the explicit topic of a text, function words are topic-independent and reflect an author's internal, subconscious syntactic habits. John Burrows demonstrated that analyzing the top 100 to 1000 MFWs via multivariate statistical matrices provides a highly robust signal for authorship attribution, as these words are immune to conscious manipulation.

Syntactic Features and POS n-grams: 2. Syntactic Features and POS n-grams: Syntactic stylometry shifts focus from individual lexical choices to the underlying grammatical architecture of prose. By tracking n-grams of Part-of-Speech tags (e.g., tracking the specific probability of an adverb preceding a preposition followed by a plural noun), models isolate the structural cadences of an author's style. This layer is especially valuable when analyzing historical texts where an author may be intentionally mimicking another writer's vocabulary, but cannot easily alter their native sentence structure.

Character-level n-grams: 3. Character-level n-grams: Character n-grams capture overlapping sequences of n characters (including spaces and punctuation marks). For example, the word 'archive' decomposed into character 3-grams yields ['arc', 'rch', 'chi', 'hiv', 'ive']. Character n-grams are incredibly resilient against OCR errors, as a single corrupted character only impacts a small subset of n-grams, leaving the broader structural profile intact. Furthermore, they subtly capture morphological prefixes, suffixes, and punctuation rhythms that human editors rarely modify.

Semantic and Vectorial Embeddings: 4. Semantic and Vectorial Embeddings: Modern computational stylometry increasingly integrates deep learning contextual word embeddings, such as Word2Vec, RoBERTa, and custom-trained transformer vectors. These high-dimensional spaces capture the contextual usage of language, mapping the semantic associations and ideological nuances unique to a specific writer's intellectual environment.

Feature Category	Linguistic Granularity	Archival Resilience	Primary Analytical Benefit
Lexical (MFWs)	Word Level (Function words, pronouns)	Moderate (Susceptible to OCR gaps)	Topic-independent; tracks subconscious writing habits.

Syntactic (POS)	Structural Level (Grammar tags, phrase trees)	High (Resilient to basic vocabulary shifts)	Isolates grammatical architecture beneath surface word choice.
Character n-grams	Sub-word Level (Overlapping character strings)	Very High (Extremely resilient to OCR errors)	Captures morphology, punctuation paths, and sub-lexical details.
Vector Embeddings	Semantic/Contextual Level (Transformer spaces)	Moderate (Requires large training corpora)	Maps ideological footprints and complex semantic networks.

3.3 Mathematical and Statistical Classification Models

Once features are extracted and normalized, they are fed into statistical and machine learning architectures to calculate similarity or execute classification. The baseline metric in computational stylometry is Burrows' Delta, an application of Manhattan distance combined with standard score normalization. Burrows' Delta measures the statistical distance between the MFW profile of an anonymous target text and the known profiles of candidate authors. The mathematical formulation is represented as follows:

$$\Delta_B = (1/n) * \Sigma_{i=1}^{n} | (f_i(T) - \mu_i) / \sigma_i |$$

Where f_i(T) represents the frequency of feature i in the anonymous text T, μ_i is the mean frequency of feature i across the entire background corpus, σ_i is the standard deviation, and n signifies the total number of features evaluated. A lower Delta value indicates a shorter statistical distance, signifying a higher probability of shared authorship.

Beyond traditional distance metrics, contemporary digital humanities heavily employs Supervised Machine Learning algorithms. Support Vector Machines (SVM) with linear or radial basis function kernels have proven exceptionally powerful for closed-class attribution problems, maximizing the geometric margin between different authors' stylistic boundaries. Random Forests and Gradient Boosted Decisions Trees offer robust alternative models, providing clear feature-importance metrics that help human scholars understand exactly which linguistic variables drove the classification decision.

In large-scale, unmapped archival ecosystems where candidate authors are unknown or incomplete, researchers pivot to Unsupervised Clustering and Authorship Verification methods. Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE) compress high-dimensional feature arrays into visually interpretable two- or three-dimensional clusters. For strict validation, the 'Impostors Method' (pioneered by Moshe Koppel and optimized by Mike Kestemont) evaluates an attribution by testing the anonymous text against the target author alongside a shifting pool of random distractor texts ('impostors'). An attribution is only validated if the target author emerges as the closest match

across a high percentage of iterations, providing a rigorous guard against false positives in sparse archival landscapes.

4. Cross-Disciplinary Case Studies in Archival Recovery

The practical power of computational stylometry is best illustrated through its deployment across diverse historical and literary contexts. This section reviews key case studies where algorithmic methodologies have fundamentally rewritten historical narratives, restored agency to erased authors, and successfully uncovered subaltern voices within complex corpora.

4.1 Recovering Nineteenth-Century Black and Indigenous Counter-Discourses

The antebellum and post-Civil War American press was a primary site for radical Black political thought and Indigenous resistance. However, due to systemic white supremacy, state-sanctioned violence, and strict economic retaliation, many African American and Native American writers were forced to publish anonymously or adopt Eurocentric, white-coded pseudonyms. Applying stylometric classifiers to digitized collections like the Black Abolitionist Papers or early tribal newspapers has allowed researchers to break through these barriers. By training models on small, verified sets of known essays by figures like Frances Ellen Watkins Harper, Maria W. Stewart, or William Apess, computational tools have successfully identified their distinct rhetorical and syntactic fingerprints within unsigned editorial columns across the historical press.

4.2 Unmasking Pseudonymity in Early Modern and Victorian Women's Writing

The history of women's literature is deeply bound to the strategic use of pseudonymity. From the Brontë sisters publishing as the Bell brothers to George Eliot hiding Mary Ann Evans' identity, women writers consistently leveraged masculine facades to secure serious critical reception. Beyond these famous examples lies a vast sea of Victorian and Early Modern periodicals where thousands of women wrote the bulk of fiction, poetry, and social critique completely anonymously. Using packages like 'Stylometry with R' (stylo), digital humanities scholars have mapped the macro-stylistic configurations of these periodicals. These computational audits have revealed that major segments of anonymous labor reform, suffrage literature, and domestic critiques were driven by interconnected networks of women writers working under collective institutional names, forcing a complete re-evaluation of the gendered dynamics of nineteenth-century print culture.

4.3 Re-attributing Subaltern Voices in Postcolonial and Transnational Corpora

In the context of colonial empires, local populations frequently utilized underground printing presses and anti-colonial broadsides to coordinate resistance and preserve local cultural knowledge. Under strict colonial surveillance, names were dangerous; texts were either completely unsigned or misleadingly attributed to fictional entities. Computational stylometry offers a powerful mechanism to bypass colonial tracking while restoring historical agency. By adapting feature extraction pipelines to handle multilingual texts and non-Western syntactic structures, transnational digital humanities projects have successfully traced the hands of specific anti-colonial intellectuals across multi-lingual archives in South Asia, North

Africa, and the Caribbean, validating the hidden intellectual labor that fueled global liberation movements.

5. Technical and Methodological Challenges

Despite its immense theoretical and reparative potential, computational stylometry is not a flawless technological oracle. When applied to marginalized archives, the methodology faces severe, structural limitations that require careful technical management and critical humility.

5.1 Dataset Sparsity and the 'Small Data' Paradox

The standard algorithms of machine learning and computational linguistics are designed for an era of 'Big Data,' typically requiring millions of tokens to train stable, highly accurate models. However, marginalized archival research operates within the exact opposite reality: a 'Small Data' paradox. Because an author's work was systematically suppressed, destroyed, or prevented from being written, the available training text for a minor writer might consist of only a few brief letters or a single short story. Attempting to train an SVM or a deep neural network on a few thousand words frequently leads to severe model overfitting, where the algorithm memorizes the specific content words of those few texts rather than isolating a stable, generalized stylistic signature. Overcoming this requires a methodological shift toward low-feature distance models, data augmentation techniques, and cross-genre normalization.

5.2 The Corrupting Influence of OCR Degradation and Textual Noise

The physical condition of marginalized archives—often printed on cheap, low-grade paper with poor ink, or preserved in damaged microfilm—results in highly corrupted digital files during OCR conversion. A stylometric model is highly sensitive to text quality. If an OCR scanner routinely turns the word 'the' into 'the' or misinterprets punctuation marks as random noise, the relative frequency matrices of MFWs and character n-grams become deeply corrupted. Studies have shown that high levels of OCR noise can artificially cluster texts based on their shared printing degradation rather than their actual authorship, creating dangerous computational illusions that can completely mislead the historical researcher.

5.3 Stylistic Mimicry, Editorial Interference, and Collaborative Heteroglossia

Style is not a static genetic marker; it is an evolving, deeply contextual, and social construct. Historical writers frequently engaged in deliberate stylistic mimicry, adopting the dominant rhetorical modes of their era to gain entry into elite literary spaces. Furthermore, historical texts rarely reached the public without heavy editing. Aggressive intervention by white male editors, publishers, and printers routinely smoothed out the unique linguistic patterns of marginalized writers, imposing a standardized, canonical style over the original prose. Finally, many radical traditions relied on collaborative authorship, where multiple individuals collectively drafted, revised, and stitched together a single document. This creates a state of heteroglossia that complicates simple single-author attribution models, requiring advanced sentence-level and section-level multi-authorship segmentation tools.

6. Ethical Frontiers: The Data Politics of Digital Resurrection

The transformation of marginalized literature into computational data introduces profound ethical questions that go far beyond standard technical validation. Scholars must critically

evaluate the socio-political implications of using data-driven tools to categorize and expose historical identities.

6.1 Algorithmic Colonialism and the Epistemic Violence of Code

Most dominant stylometric tools and NLP models were built, tested, and optimized on large Western, Anglophone, and canonical datasets. When these identical tools are unthinkingly applied to subaltern, non-Western, or vernacular texts, they can perpetrate a form of algorithmic colonialism. Imposing rigid, Eurocentric models of grammar, structural syntax, and standardized textuality onto non-standard dialects or oral-influenced prose risks erasing the unique linguistic innovations of marginalized cultures. Digital humanities must ensure that code bases are ethically re-engineered to respect the cultural specificity and historical context of the literature they analyze.

6.2 The Ethics of Unmasking: Consent, Safety, and Historical Agency

Anonymity was frequently a conscious, protective choice made by marginalized authors to guarantee survival, preserve personal privacy, or protect families from violent reprisal. Using computational tools to forcefully strip away an author's historic pseudonym is an exercise of immense interpretive power. Scholars must pause to consider: does forcing a historically anonymous author into the light of contemporary classification respect their original agency, or does it violate their historical right to opacity? The drive for total archival transparency must be balanced against an ethics of care, carefully weighing the historical benefit of recovery against the potential violation of past protective silences.

6.3 Collaborative Workflows: Integrating Computational Outputs with Critical Humanism

To prevent computational stylometry from descending into a cold, hyper-positivist numbers game, digital humanities must champion a collaborative workflow that pairs algorithmic output with traditional critical humanism. An algorithmic attribution score or a high-probability clustering map should never be treated as an absolute, unassailable statement of historical fact. Instead, computational outputs must serve as interpretive prompts—statistical pointers that guide the human archivist, cultural historian, and literary scholar toward deep close reading, comprehensive biographical contextualization, and rigorous cross-verification with physical evidence. The machine provides the coordinates; the human scholar provides the understanding.

7. Conclusion and Future Horizons

Computational stylometry represents a major paradigm shift for archival recovery and digital humanities scholarship. By transforming texts into quantitative, multi-dimensional structures, these tools allow us to bypass the physical omissions and cataloging biases of traditional history, offering a robust method to detect the subconscious linguistic fingerprints of authors whose identities were stripped away by systemic oppression.

As we look to the future, the integration of large language model embeddings, advanced multi-author segmentation, and error-resilient character architectures will continue to enhance the accuracy and scope of archival recovery. However, the true success of this movement will

not be measured by algorithmic complexity or classification accuracy alone. It will be measured by the degree to which these tools are grounded in ethical reflection, cultural specificity, and a commitment to historical justice. By deploying computational stylometry to recover lost voices, digital humanities can move beyond simple canonical curation, fundamentally transforming the archive from a monument of historical power into an open, dynamic space of democratic memory.

References

Agarwal, D. B. (2026). Artificial Intelligence in Language and Literature. International Journal of English and Studies (IJOES), 8(2), 413-422.

Burrows, J. (2002). 'Delta': a Measure of Stylistic Difference and a Guide to Authorship Attribution. Literary and Linguistic Computing, 17(3), 267–287. https://doi.org/10.1093/llc/17.3.267

Derrida, J. (1996). Archive Fever: A Freudian Impression. University of Chicago Press.

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. The R Journal, 8(1), 107-121. https://doi.org/10.32614/rj-2016-007

Foucault, M. (1972). The Archaeology of Knowledge. Routledge.

Hollingsworth, C. D. (2012). Syntactic Stylometry: Using Sentence Structure for Authorship Attribution (Master's thesis, University of Georgia).

Huang, B., Chen, C., & Shu, K. (2025). Authorship Attribution in the Era of LLMs: Problems, Methodologies, and Challenges. ACM SIGKDD Explorations Newsletter, 26(1), 21-43. https://doi.org/10.1145/3715073.3715076

Jockers, M. L. (2013). Macroanalysis: Digital Methods and Literary History. University of Illinois Press.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., & Daelemans, W. (2016). Authenticating the Writings of Julius Caesar. Expert Systems with Applications, 63, 86–96. https://doi.org/10.1016/j.eswa.2016.06.029

Koppel, M., Schler, J., & Argamon, S. (2009). Computational Methods in Authorship Attribution. Journal of the American Society for Information Science and Technology, 60(1), 9-26. https://doi.org/10.1002/asi.20961

Michailidis, P. D. (2022). A Scientometric Study of the Stylometric Research Field. Informatics, 9(3), 60. https://doi.org/10.3390/informatics9030060

Moretti, F. (2013). Distant Reading. Verso Books.

Stamatatos, E. (2009). A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556. https://doi.org/10.1002/asi.21001

Trouillot, M. R. (1995). Silencing the Past: Power and the Production of History. Beacon Press.

Unguendoli, F., Cristadoro, G., & Beghelli, M. (2018). Stylometry and Automatic Attribution of Medieval Liturgical Monodies. Italian Journal of Computational Linguistics, 4(1), 91-105. https://doi.org/10.4000/ijcol.518

van Noord, N. (2022). A Survey of Computational Methods for Iconic Image Analysis. Digital Scholarship in the Humanities, 37(4), 1316-1338. https://doi.org/10.1093/llc/fqac003